A general guide to data preparation for computer analysis of farm survey data

Hesse, E.

1982

A GENERAL GUIDE TO DATA PREPARATION FOR COMPUTER ANALYSIS OF FARM SURVEY DATA

Edith Hesse de Polanco*

Economics Training Note
1982

CENTRO INTERNACIONAL DE MEJORAMIENTO DE MAIZ Y TRIGO
INTERNATIONAL MAIZE AND WHEAT IMPROVEMENT CENTER
México

A GENERAL GUIDE TO DATA PREPARATION FOR COMPUTER ANALYSIS OF FARM SURVEY DATA

Edith Hesse de Polanco*

Economics Training Note
1982

* Research Assistant, CIMMYT Economics Program, Mexico. The views expressed are not necessarily those of CIMMYT.

January, 1982

Training Note

A General Guide to Data Preparation for Computer Analysis of Farm Survey Data

Edith Hesse de Polanco

1.0 Introduction

If a computer and an appropriate program for survey data analysis are available, the researcher has to decide whether or not these computer facilities should be used to analyse the survey data or whether an analysis by hand is sufficient. In almost all cases a preliminary analysis by hand is very useful to get a "feel" for the data. In our experience a complete manual analysis will be more efficient if the number of farmers in the sample is less than 50. If the sample size is about 50-100 farmers, a computer analysis may be helpful when the study area is rather complex and farmers' practices and circumstances need to be analised through a relatively large number of cross-tabulations. Surveys with more than 100 farmers can nearly always be analysed more efficiently by a computer if appropriate computational facilities are available.

This note provides a short overview on presently used computers ("hardware") and computer programs ("software") assuming that the researcher has no previous experience with computers. It also presents the steps involved in preparing the data for computer analysis.

2.0 Hardware

There are effectively three types of computers: microcomputers, minicomputers and the large "mainframe" computers. The differences most noticed between these three computer types lies in their speed of computation and in the amount of memory. The computer memory is usually expressed in numbers of K bytes[1/]. Any computer has a main or central part consisting of the power supply and the central processing unit (CPU). Additionally there may be devices such as video displays with keyboards, card readers, disk and tape drives, printers and other peripheral devices.

Microcomputers often are not larger than a normal typewriter. They usually have an in-built video display and their memory generally does not exceed 64 K bytes. Data are typed in through the keyboard and may be stored on a floppy disk or on a cassette. As microcomputers are a relatively recent phenomenon, a shortage of appropriate software still exists. However it is expected that these computers will be increasingly used in agricultural research in developing countries primarily because of their relatively low costs[2/].

Minicomputers normally have a memory larger than 64 K bytes. They may be connected to larger "mainframe" computers and data are usually read in through a card reader. Data can later be stored on

---

1/ All computers use the binary system, i.e. any number, letter or special character is expressed by a combination of zeros or ones. Each 0 or 1 is called a "bit". The minimum amount of bits necessary to represent a character is called a "byte". Finally, 1024 bytes form one "K bytes".

2/ A 48 K byte microcomputer system with a floppy disk drive and a printer presently costs about US $ 3000-4000.

tapes or disks. Minicomputers normally use computer programs similar to those of large "mainframe" computers, although the prepared statistical packages (described below), usually are too large for minicomputers.

Large "mainframe" computers such as the IBM 360 and the IBM 370 have a much larger memory[1/] and work at a considerably higher speed. Apart from their wide use in developed countries they are now installed in many research and governmental institutions in developing countries. However, in some cases the appropriate software is not available or has not been successfully implemented, so that some of these computers are not being used at their full potential.

The above comparison between mini and micro, and micro and mainframe computers can serve as a practical guide. However, the rapid changes in computer technology are making the described differences less recognizable.

3.0 Software

The crucial factor in the usefulness of a given computer for the researcher is the availability of software. There are a number of programs and software packages which may be used for the analysis of farm survey data. Some are only available on large computers, others are designed for minicomputers - but can also be used on large computers - and very few are presently available for microcomputers (although this situation is changing rapidly). For example, a small program for the analysis of experimental data, including economic analysis, will soon be available for a microcomputer[2/]. Two other small programs specially

---

1/ This can range from 256 K bytes up to several Megabytes (1 Megabyte = 1024 K bytes).

2/ Stilwell, T.C., Manual del usuario del sistema de estadística agrícola - Para el digital PDP 11/45, Consortium for International Development, Cochabamba - Bolivia.

designed for the analysis of farm survey data are presently available for mini-computers[1/].

However, if a large computer with one of the more widely used statistical packages is available, the researcher should try to get access to it especially for the analysis of larger surveys. The Statistical Analysis System (SAS) allows the analysis of experimental data as well as of survey data. The Statistical Package for the Social Sciences (SPSS) also has many facilities for the analysis of survey data. Both packages consist of a large number of statistical procedures and are highly flexible with respect to data manipulation and representation. An inexperienced researcher needs only a day or so to get acquainted with the basic instructions, contained in the special user's guides[2/] although some guidance from the local computer staff may be useful!

4.0 Data Preparation

Data preparation begins before the questionnaires arrive in the office. The researcher has to check and edit every questionnaire thoroughly as soon as possible. At this stage inconsistencies should be

---

1/ FAO, FARMAP, User's Manual, Farm Management Data Analysis Package, Rome, 1981.
Hesse de Polanco, E. & P. Walker, A User's Guide to FASAP, A FORTRAN Program for the Analysis of Farm Survey Data, CIMMYT Economics Working Paper, Sept. 1980. The latter program has been developed for a minicomputer but is also readily usable in larger computers. It has facilities for data transformation and missing values, performs one way frequencies, cross-tabulations and tables of means by group, all with the associated statistics.

2/ SAS Institute, SAS Introductory Guide, 1978, SAS Circle Box 8000, Cary, North Carolina 27511. SAS Institute, SAS User's Guide, 1979 Edition. Post Office Box 10066, Raleigh, North Carolina 27605. Both guides are available at US$ 10.00 at their respective addresses.
NIE, N.H., et. al., Statistical Package for the Social Sciences, McGraw Hill, 1970, 1975.

cleared up while the enumerators have the interviews still fresh in mind. Unlikely values or illegible data may be noted and sometimes a revisit is required. In the case of serious problems, certain questionnaires might be discarded.

In cases where the sample size is rather small (less than 50 farmers) values for the most important variables are written by hand for each farmer onto a large sheet of paper. This facilitates the manual calculation of simple frequency distributions and means. If these calculations are to be performed for specific groups of farmers, the task becomes increasingly time consuming using manual analysis.

Once the decision to use a computer for data analysis has been made, the data have to be coded. This means that all important information from the questionnaire has to be transferred onto coding sheets[1/] and later onto punch cards according to precise rules. Every variable has to be identified by a variable name or number and code categories for each variable have to be determined. This is usually done by preparing a so called "code-book". The beginning of a typical code book is shown in Table 1. It contains three major pieces of information: 1) a number and/or a shortened name for each variable of the questionnaire; 2) the code categories for each variable; and 3) the card number and column range into which codes for these variables have to be punched onto cards.

1/ A typical coding sheet is included in the annex. Each line of a coding sheet is later punched onto one card. The above mentioned preliminary manual analysis might also be done from the coding sheets or from a later computer listing of the data.

Table 1. CODE BOOK

Variable Names	Code	Column Range
Identification[1/]		
VILLAGE	1= Tequesquinahuac 2= Huexotla 3= Tlaixpan	1
FARM	No. 1-100	2-4
CARD	Card Number =1	5
V1	Number of plots	6-8
V2	Hectares of wheat	9-11
V3	Hectares of maize	12-14
V4	1= flat 2= some slope 3= steep	15-17
.		
.		
.		
V25	1= tractor use-yes 0= tractor use-no	78-80
VILLAGE	the same as in card number 1	1
FARM		2-4
CARD	Card Number =2	5
V26	Tractor Rental $/ha	6-8

1/ See paragraph 3.2 Identification Code.

Continuation Table 1..

V27	1= not enough moisture 2= tractor not available 3= didn't have time 4= not enough moisture and not enough time.	9-11
V28	1= owned 2= rented 3= community 4= government	12-14
V29	1= fertilizer use-yes 2= fertilizer use-no	18-20
.		.
.		.
.		.
V50	lt of water per ha used for herbicide application	78-80

VILLAGE	the same as in card number 1	1
FARM		2-4
CARD	Card Number = 3	5
V51	1= manual 2= animal 3= tractor	6-8
.		.
.		.
.		.

etc.

4.1 Variable Names

In most computer programs, variable names should begin with a letter, should not be longer than 8 characters and should not include blank spaces between the characters. It is normally a good idea to identify those variables to be coded by consecutive numbers (i.e. V1, V2,...). In certain cases it might be useful to identify certain variables by partial names instead of numbers, especially for those variables that are analysed frequently. For example, if topography is an important variable for cross-tabulation, the variable V4 in Table 1 might be better labeled with "TØPØGR". In the same way, if a number of cross-tabulations by village should be done, "VILLAGE" could be used as a variable name. However, it is impractical to choose partial names for the bulk of the variables.

4.2 Code Categories

In general, codes should be numbers although some statistical packages allow the use of letters or special characters.

Some questions which often arise when appropriate code categories are to be chosen are discussed below:

1) Quantitative or continuous data (e.g. number of hectares of a given crop) should always be coded as actual numbers. One should never categorize numerical data as this can be done much easier afterwards by data transformations in the program. For example, never code area as 1 = 0-10 ha, 2 = 11-20 ha, etc., if actual area is known. This implies a loss in available information and flexibility.

2) Qualitative or discrete data can be coded by assigning a number to each category. For example, seed source might be coded as follows:

1 = own seed

2 = from a neighbor

3 = from the bank

4 = other _____

It is often sufficient to use a separate code only for the most commonly occurring categories and to code all residual observations as "other".

3) Subjective data, e.g. opinions and qualitative data, should be grouped into similar categories.

Example: Why didn't you plough in November?

a) Ground was too hard

b) Not enough moisture

c) Couldn't obtain tractor

d) Tractor has been out of working order

e) Off-farm work

f) Busy in other farm work

g) Other _____

In such an "open" question any number of subjective reasons might appear in the questionnaires. In the process of editing it is usually convenient to group togehter some of the answers in

order to end up with a reasonable number of categories. However, in certain cases one farmer might mention two or more categorized reasons to the same question. In such a case the researcher has to decide if it is worthwhile to introduce an additional code which indicates the combination of two categories (e.g. Code 4 = not enough moisture and off-farm work).

4) For coding dates, e.g. weeks or month in which a given practice has been done, it is often best to use the number of days or weeks from a key reference point.

Example:
0 = harvest month of the previous crop
1 = one month after harvest
2 = two months after havest
.
.
.
12 = etc.

In this case the time range from one reference point to an other may easily be calculated in the computer program.

5) If the farmer is asked to indicate the quantity of a given input, it is always important to ask and code first a so called "yes/no" question.

Example:
1. Did you use herbicide? yes/no
2. How much did you apply? lt/ha_____

If the first question wouldn't have been included, the temptation to put a zero for a "non-user" into the data field for question No. 2 becomes evident. For "non-users" question No. 2 has to be coded with a missing value indicator[1/].

1/ See paragraph 3.5 for the description of missing value indicators.

The same occurs with the performance of certain practices with their related questions:

Example:
1. Did you plough?
2. Date of ploughing?
3. Implement for ploughing?

etc.

If the first question is answered by "no" all following related questions should be coded with a missing value indicator.

6) The coding of fertilizer data may cause problems since there often exist a number of N and P products and compounds. The best way to handle it is to manually calculate nutrients applied and to code then these nutrient quantities. A possible exception is when the form of fertilizer is itself a variable.

7) Code categories themselves may be categorized to facilitate the analysis and data interpretation. For example, barley varieties might be coded as follows:

Example:

11 = Cerro Prieto	
12 = Puebla	New varieties
13 = Centinela	
24 = Apizaco	
25 = Porvenir	Old varieties
26 = Chevalier	

In this example a Code 1 and 2 were chosen to identify the two variety groups "new" and "old" respectively. Additional codes identify each variable in the two groups. Interpretation of frequency tables is easier if coding is organized this way.

8) The researcher always must have in mind what type of analysis is to be done with each variable in order to determine the correct form for coding, i.e. coding several variables from one question or coding only one variable with several code categories from the same question.

Example: In an irrigated area in northern Mexico where two crops by season were grown, researchers were interested in knowing to what extent the weed problems observed in the field were related to the preceding crops. In this case the farmer was asked: "Which crop did you plant on this field in:

1980	1979	1978
Summer/Winter	Summer/Winter	Summer/Winter

There was an initial temptation to code this question using six different variables, i.e. one for each crop cycle. But as the required analysis was a crosstabulation of "weed problems" by "previous rotation", it was decided to code it in the following way: Only one variable called "previous crop" was coded, using the following categories:

1 = Cotton 1980
2 = Safflower 1980
3 = Other row crop 1980
4 = One year continuous wheat
5 = Two years continuous wheat
6 = Three years continuous wheat

9) Code categories should be uniform not only within one survey but also from one survey to the other in order to make the coding task more straightforward. For example, one should always use a 0 for "no" and a 1 for "yes", or a 1 for "manual", a 2 for "animal" and a 3 for "tractor", etc.

4.3 Column Range

In our experience we found it useful to assign the same number of columns to each coded variable[1/]. This means that even in the case of a yes/no question, the data field should be three columns wide, even though the code for yes (usually a 1) and the code for no (usually a 0) will only occupy one column. This code enters into the right justified column of the data field, leaving the left two columns blank. In very few cases, where certain quantities might occupy a four column data field, e.g. tractor rental: 1800 $/ha the values for every farmer of this tractor rental variable should be coded by dividing all values by 10.

4.4 Identification Code

A complete identification code allows the researcher to identify each data card in a unique manner. In cases where a two stage sampling of farmers is used (e.g. a village sample and then a farmer sample), the identification normally will consist of two different codes: one referring to the village and the other referring to the farmer[2/]. In most cases the coding of all important variables

---

1/ The column range assigned to a certain variable is usually called a "data field".

2/ In the computer language you will later use the expression "observation", "case" or "unit" instead of "farmer".

from one questionnaire will occupy more than one punch card (i.e. more than 80 columns). In these cases the card number also has to be coded (see Table 1). Some people even use an identification code for the survey, e.g. they put a special code for survey in the beginning or final columns of each data card.

4.5 Missing Value Indicators

At the beginning of any computer analysis it is important to determine how missing data are to be handled. In survey data two types of missing data are usually found. In some cases the farmer uses a certain input, but does not remember the quantity or the date when he applied it. In very few cases the farmer may amply decline to answer a certain question. For these cases a missing value indicator for "no response"[1/] should be coded. The second type of missing value indicator is used in those cases where a certain question is not appropriate to the specific situation of the farmer. For example, it is senseless to ask the farmer whether he used an owned or rented tractor, if we know from a previous question that he used no tractor at all. A missing value indicator for "not appropriate question"[2/] should then be coded into the data field of the variable "tractor ownership". The form and the handling of missing value indicators depend on the software. Therefore it is important to know which program or system package will be used in order to observe the existing rules with respect to the missing values before data are coded.

1/ Using SAS for our data analysis we coded this type of missing values as "R" and using FASAP we coded it as "-1".

2/ Using SAS we coded it as "N" and using FASAP as "-2".

4.6 Additional Hints for Data Preparation

In the process of coding the order of questions (i.e. variables) should not be changed. That is, each variable should be coded in the same order as it appears in the questionnaire. This doesn't mean that clearly worthless variables should not be omitted. For example, the variable tractor use - yes/no should not be coded if all sampled farmers were using a tractor and in the same way the variable herbicide use - yes/no should not be coded if no sampled farmer used a herbicide[1/].

In our experience the best way to transfer the data from the questionnaires onto the coding sheets is the following: All variables from one questionnaire should be coded in one step, using different coding sheets if necessary. For example, if 25 variables can be coded into the 80 columns of one line of the coding sheet (i.e. the 80 columns of one punch card) it is best to use one coding sheet for the first 25 variables and the following set of 25 variables is coded onto a second coding sheet; the next 25 variables onto a third coding sheet and so on - repeating the identification code and identifying each coding sheet by a consecutive number (see example below and also Table 1) . When all variables of the first questionnaire have been coded into the first line(s) of the coding sheet(s), then all variables from the second questionnaire will be coded into the second line(s) of the coding sheet(s) occupying the same data fields for each variable.

1/ However this will rarely occur when the questions in the questionnaire are based on the information obtained by a good exploratory survey.

Example:

Village	Farm	Card	V1	V2	..	.	V25
1	1	1	2	20	. . .		1
1	2	1	3	35	. . .		0
.	.	.	.	.			.
.	.	.	.	.			.
.	.	.	.	.			.
1	25	1	2	60	. . .		1

Village	Farm	Card	V26	V27	..	V50
1	1	2	300	1	. . .	200
1	2	2	-2	4	. . .	-1
.	.	.	.	.		.
.	.	.	.	.		.
.	.	.	.	.		.
1	25	2	350	2	. . .	200

Village	Farm	Card	V51	V52	.	.	.	V75
1	1	3	1	0	.	.	.	-1
1	2	3	2	1	.	.	.	2
.	.	.	.	.				.
.	.	.	.	.				.
.	.	.	.	.				.
1	25	3	2	0	.	.	.	3

This type of coding has the following advantages:

1) Coding all variables from one questionnaire in one step allows detection of further inconsistencies in responses.

2) After coding has been finished, each data field may be easily checked in a vertical manner and values which do not correspond to the category range established for each data field may be detected and corrected.

5.0 Key-punching

After finishing the task of transferring the data onto coding sheets[1/], the data usually will be punched onto punch cards[2/].

1/ It is not always necessary to transfer the data onto coding sheets since the preparation of a precoded questionnaire allows key-punching right from the questionnaire. However, such a precoded questionnaire is less flexible, e.g. does not allow omission of certain irrelevant variables from coding or does not allow "open questions". Enumerators also may have more difficulties in handling the codes and a precoded questionnaire usually becomes more voluminous. Inconsistencies which often are detected in the process of coding and which can still be cleared up cannot be detected by a key-puncher.

2/ In the case of microcomputers data usually are directly typed on keyboard and stored on a floppy disk or a cassette.

Key-punching usually is done by a specially trained key-punch operator and is verified by redoing it on a verification machine. In certain cases cards may have to be interspersed after punching. A data printout should be requested immediately afterwards to allow the researcher himself a thorough check of his coded data. Errors will be marked on the punch-cards and corrections often may be done by the researcher himself.

After data checking and correction data usually will be put onto a disk or tape file because punch-cards easily become damaged if they are put into the card reader many times. It is also more expensive to read the data from the cards instead of reading them from a tape or disk file since a unit cost per card read will be charged. However, a temporarily file on tape or disk might be deleted accidentally so that cards always should be stored in a safe place considering that humid conditions may cause deterioration.

Program instructions are generally written also onto the coding sheets and later punched onto punch cards. Some simple rules should be observed in order to prevent mistakes during key-punching:

- the letter O should be distinguished from the number zero by crossing the letter O by a slash (Ø).

the letter I has to be distinguished from the number 1 by putting two horizontal bars onto the letter I.

- the letter Z should be written with one horizontal bar in order to distinguish it from the number 2.

- the letter G should be carefully distinguished from the number 6.

CENTRO INTERNACIONAL DE MEJORAMIENTO DE MAIZ Y TRIGO

PROGRAMA _____

HOJA DE CODIFICACION

FECHA _____
HOJA _____ DE _____
PROGRAMO _____

5 10 15 20 25 30 35 40 45 50 55 60 65 70 75